

利用「R」分析漢字——邁向漢字統計學的建立

林偉業

香港大學教育學院 jwilam@hku.hk

本研究旨在以開源數據統計分析軟體「R」建立漢字分析的資料庫，並利用資料庫作統計分析，模擬漢字字音的歷史演變，作為漢字統計學的應用示例。

R 是近年流行與統計學界的開源軟體，長於數據統計分析，以及數據的視覺化表達。本論文的報告的重點有二：

一、利用 R 的數據結構建立漢字的關係資料庫：漢字多義多音、但音義關係不必然對應，造成建立漢字雙向列聯表 (Two-Way Contingency Table) 式資料庫的困難，本研究嘗試以 R 解決這一難題；

二、利用 R 的數據分析，計算字音演變規律，解決香港社會上有關正音的爭議。

漢字統計學是漢字研究與教學必不可少的基礎。本研究希望藉以上兩項，為建構更完整的漢字統計學，提出更清晰的方向。

Towards quantitative analysis of Chinese character

- using "R" to analyze Chinese character

LAM Wai Ip, Joseph

Faculty of Education, The University of Hong Kong

jwilam@hku.hk

This study aims at building database of Chinese character by using an open source statistical software “R”, and simulating the historical change of pronunciation of Chinese character by using “R”.

“R” has becoming a more and more popular statistical software. It is good at statistical analysis and visualizing data. The paper is going to discuss the following two issues:

1. How to use “R” to build relational database for Chinese character to solve the difficulty in building database in 2-way contingency table due to the irregular mapping of sound and meaning of the polyphonic and polysemous Chinese character.
2. How to use “R” to analyze Chinese to develop pronunciation rule of Chinese character to solve the argument on "correct pronunciation" in Hong Kong.

Quantitative analysis of Chinese character is one of founding stones of researching and teaching Chinese character. How to establish such discipline will also be discussed in this study.